Chapter 2 - Descriptive Statistics

2.1 Describing Numerical Data-
Dot Plots, Histograms, Stem and Leaf Diagrams

2.2 Measuring the Centre
Mean, Median, Mode

2.3Measuring Variability
Range, Variance, Standard Deviation

2.4 Relative Standing (How do you compare?)
Percentile, Z-Score

2.5 Quartiles and Box Plots
Quartiles, IQR, Box-Plots

Section 2.1 Describing Numerical Data-

Dot Plots, Histograms, Stem and Leaf Diagrams

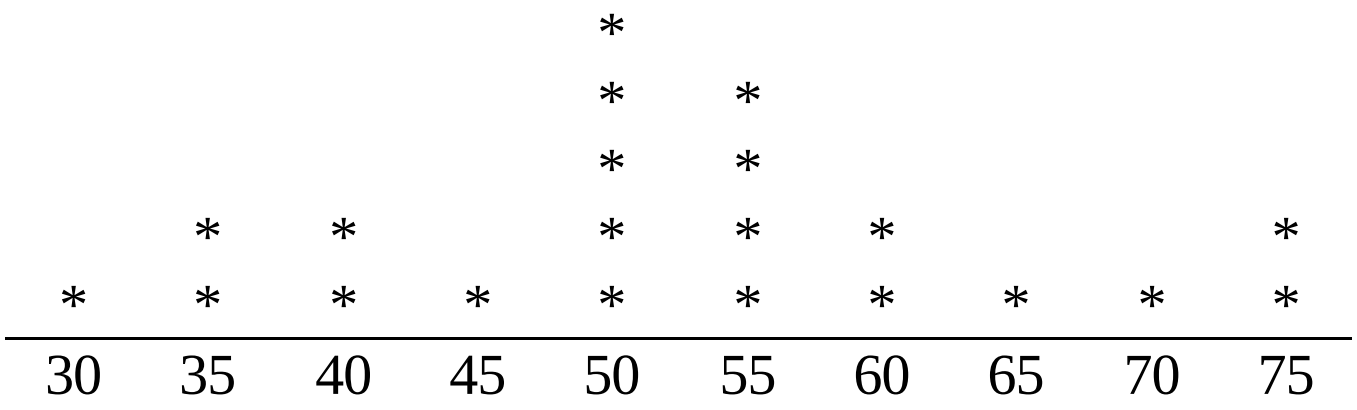
2.1.0 Dot Plots

A long time ago before computers could draw decent graphs they used to draw Dot Plots as a way of describing numerical data. The range of values in the dataset are marked on an X-axis and then a dot is placed above the relevant point on the axis for each value in the dataset. If two or more observations have the same value then dots are stacked on top of each other. You should use minitab to generate some Dot Plots

*** 2.1.1 Example

Draw a Dot Plot for the following dataset

50 35 70 55 50 30 40 65 50 75 60 45 35 75 60 55 55 50 40 55 50



2.1.2 Stem and Leaf Diagrams

Stem and Leaf Diagrams are graphical ways to display a group of integers in a dataset.

Steps for Constructing a Stem and Leaf Diagram

1. Select one or more of the leading digits to be the Stem values, the remaining digits become the Leaves.
2. List Possible Stem values in a column
3. Record the Leaf for every observation beside the corresponding Stem value.
4. Indicate on the display what units are used for the Stems and Leaves.

2.1.3 Example

Measurements are taken of the molar polarization of gaseous water at 100kPa. Some of these measurements are given below in units of cm^3mol^{-1} .

71 52 52 75 64 60 48 56
67 29 11 53 25 46 58 46
49 62 66 40 19 54 57 54
60 19 59 43 51 40 21 45
46 62 73 59 36 45 55 46
45 32 55 46 51 46 65 49 61 40

A Stem And Leaf Diagram will look like this:

1	1 9 9
2	1 5 9
3	2 6
4	0 0 0 3 5 5 5 6 6 6 6 6 6 8 9 9
5	1 1 2 2 3 4 4 5 5 6 7 8 9 9
6	0 0 1 2 2 4 5 6 7
7	1 3 5

STEM UNIT = TENS
LEAF UNIT = ONES

2.1.4 Frequency Distributions for Continuous Data

A meaningless frequency distribution:

Dataset: 12.0 12.3 13.1 14.2 11.5 12.7

Value	Frequency
11.5	1
12.0	1
12.3	1
12.7	1
13.1	1
14.2	1

It makes much more sense when dealing with continuous numerical data to define **Class Intervals** on the REAL line which may contain several observations which are close together if not exactly the same.

2.1.4B Example
A more Meaningful Frequency Distribution

Using the same dataset as before:
Dataset: 12.0 12.3 13.1 14.2 11.5 12.7

This time we split up the observations into 4 intervals instead of looking at individual values.

These Class Intervals are:

[11.0, 12.0) [12.0, 13.0) [13.0, 14.0) [14.0, 15.0)

which we could also write as
11.0-<12.0 12.0-<13.0 13.0-<14.0 14.0-<15.0

So the Frequency Distribution looks like this:

Class Interval	**Frequency**
11.0-<12.0	1
12.0-<13.0	3
13.0-<14.0	1
14.0-<15.0	1

NOTE:
The value 12.0 belongs to the interval 12.0-<13.0 and not to the interval 11.0-<12.0.

2.1.5 Relative Frequency Distributions for Continuous Data

We define the relative frequency of a particular Class Interval to be the Frequency of that Interval divided by the Total number of observations in the dataset.

2.1.5A Example

The compound 1,3,5-trichloro-2,4,6-triflourobenzene is an intermediate in the conversion of hexachlorobenzene to hexaflourobenzene. We examine its thermodynamic properties by measuring it's heat capacity over a range of temperatures. 40 measurements of the heat capacity ($JK^{-1}mol^{-1}$) taken at temperature=15^0K are presented below.

8.0 12.9 13.0 8.9 10.1 7.3 11.1 10.9 6.2 8.1 8.8 10.4 15.7 13.6 19.3 9.9 8.5 11.1 10.7 8.8 10.7 6.8 7,4 4.8 11.8 13.0 9.5 8.1 6.9 11.5 11.2 13.6 4.9 18.8 15.7 10.8 10.7 11.5 16.1 9.9

We will define class intervals as:
3-<6 6-<9 9-<12 12-<15 15-<18 18-<21

This gives us the following Frequency/Relative Frequency Distribution

Class Interval	Frequency	Relative Frequency
3-<6	2	0.05
6-<9	12	0.3
9-<12	16	0.4
12-<15	5	0.125
15-<18	3	0.075
18-<21	2	0.05

Looking at this table we can now answer some questions, such as:

A. How many measurements have a heat capacity of less than 12 $JK^{-1}mol^{-1}$?
B. What proportion of measurements have a heat capacity of between 9 and 15 $JK^{-1}mol^{-1}$?

Answers:
A. 2 + 12 +16 = 30
B. 0.4 + 0.125 = 0.525

2.1.6 Cumulative Relative Frequencies

Cumulative Relative Frequencies measure the proportion of observations falling below a specified value. The Cumulative Relative Frequency for a particular Class Interval is calculated by summing up the Relative Frequencies for that Class Interval together with the Relative Frequencies for all previous Class Intervals

2.1.6A Example

The following shows CRFs for the heat capacity example

Class Interval	Relative Freq	Cum Rel Freq
3-<6	0.05	0.05
6-<9	0.3	0.05+0.3 = 0.35
9-<12	0.4	0.05+0.3+0.4 = 0.35 + 0.4 = 0.75
12-<15	0.125	0.75 + 0.125 = 0.875
15-<18	0.075	0.875 + 0.075 = 0.95
18-<21	0.05	0.95 + 0.05 = 1.0

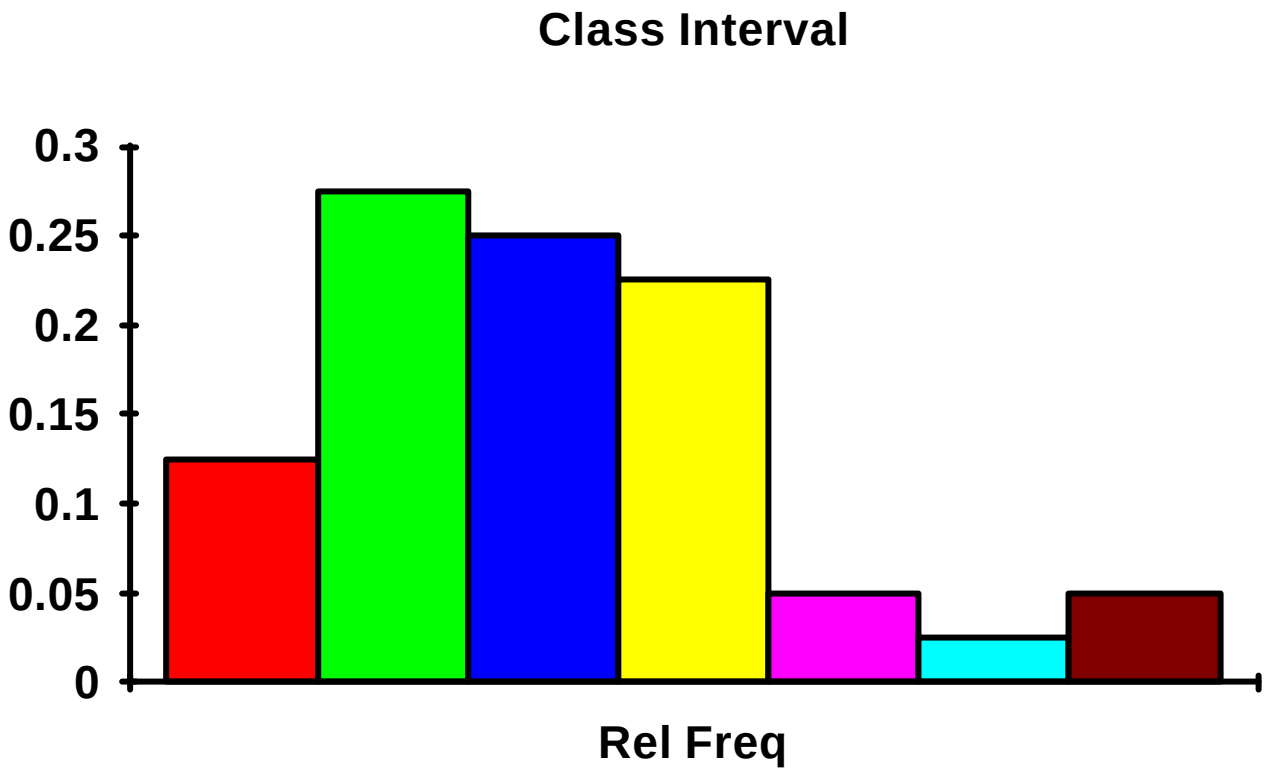
2.1.7 Histograms for Continuous Data when Class Intervals have Equal Width

When we constructed Frequency/Relative Frequency Distributions for Continuous Data so far the Class Intervals we used all had the same widths. When this is the case it is easy to go one step further and draw Histograms for such data. These Histograms are drawn in almost exatly the same way as all the previous ones.

1. Mark the ends of each Class Interval on the horizontal X-axis
2. Mark either the Frequencies or Rel Freq.s on the vertiucal Y-axis
3. Draw a rectangle for each class directly above the corresponding interval so that the sides of the rectangles are at the ends of the Class Intervals which you have marked on the X-Axis.

2.1.7A Example

Class Interval	**Freq**	**Rel Freq**
0-<10	5	0.125
10-<20	11	0.275
20-<30	10	0.250
30-<40	9	0.225
40-<50	2	0.05
50-<60	1	0.025
60-<70	2	0.05
Total	**40**	**1.0**



2.1.8 Histograms of Continuous Data when the Class Interval Widths are Unequal

In all of the previous examples of Histograms the height of each rectangle represented the Frequency or Relative Frequency. Now the area of each rectangle is just the base multiplied by the height. Since all rectangles had the same base we find that the area of each rectangle was proportional to the Frequency or Relative Frequency.

It is not always neccessary to have the bases of the rectangles be the same width, in many cases it makes more sense to have the bases of some rectangles be wider than others. If we do this however we must change the way that we draw Histograms.

From now on we are going to draw Histograms so that the AREA of each rectangle represents the Relative Rrequency.

If the area is the Relative frequency we get the following relationship:

Relative Frequency = Area = Base * Height

So we find

Height= (Relative Frequency)/Base

This new Height measurement we will call Density and since the Base of each rectangle is actually the width of a Class Interval we end up with a final equation which we will use from now on.

Rectangle Height =
Density =
(Relative Frequency)/(Class Interval Width)

2.1.9 Example

Class Interval	Relative Frequency	CI Width	Density
-50% -< -10%	0.023	40	0.000575
-10% -< -5%	0.055	5	0.011
-5% -< -2.5%	0.097	2.5	0.0388
-2.5% -< 0%	0.21	2.5	0.084
0 -< 2.5%	0.189	2.5	0.0756
2.5% -< 5%	0.139	2.5	0.0556
5% -< 10%	0.116	5	0.0232
10% -< 50%	0.171	40	0.004275

INSERT EXAMPLES of Correct/Incorrect Histogram for this data

Example 2.1.10 Marijuana Usage
A telephone survey was conducted on marijuana usage. The frequency distribution gives the amount of marijuana in grams smoked per week for those drug fiends **(respondents)** who indicated they used the drug.

Amt Smoked	**Frequency**	**Rel Freq**	**Cum Rel Fr**
0 -< 3	94	0.188	0.188
3 -< 11	269	0.538	0.726
11 -< 18	70	0.14	0.866
18 -< 25	48	0.096	0.962
25 -< 46	10	0.02	0.982
46 -< 60	7	0.014	0.996
60 -< 74	2	0.004	1
Total	500	1	

Calculate the Relative Frequencies & CRF for this dataset.

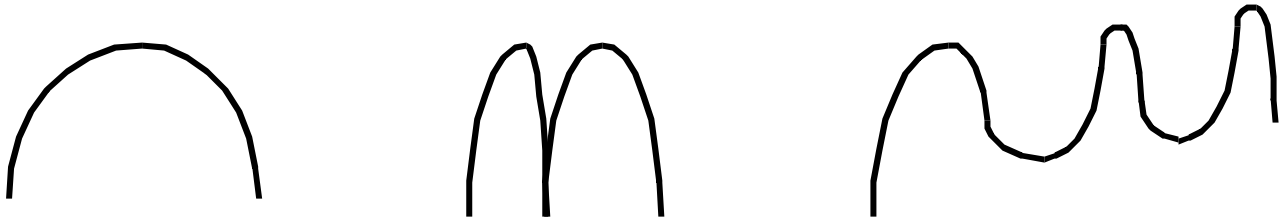
Draw a Histogram of this data.

What proportion of respondents smoked less than 25g per week?

Approximately what proportion of respondents smoked more than 53g per week?

2.2.18 A final word on Histograms

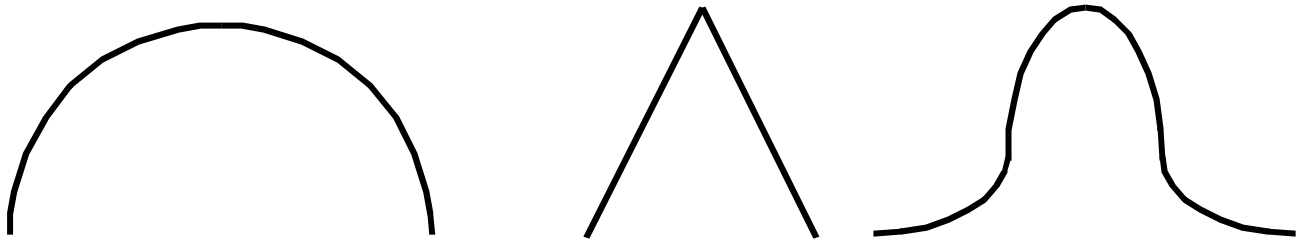
The general shape of a histogram is important.
The number of peaks in the histogram determines whether a distribution is classed as Unimodal, Bimodal or Multimodal.



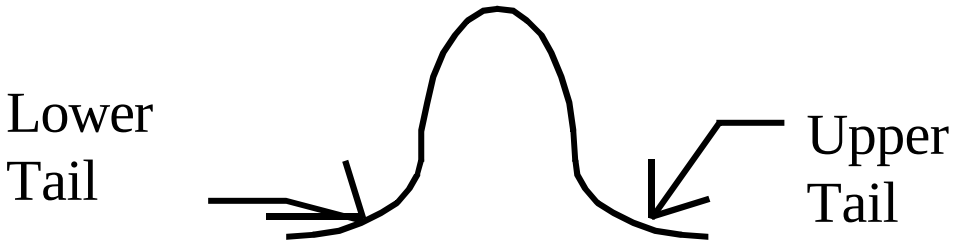
UniModal BiModal MultiModal

In addition to this classification we can further classify UniModal distributions as to whether they are symmetric or not.
A unimodal distribution is defined to be Symmetric if there is a vertical line of symmetry through the middle of the distribution such that the distribution to the left of this line is the mirror image of the distribution to the right of this line.

These 3 distributions are symmetric:

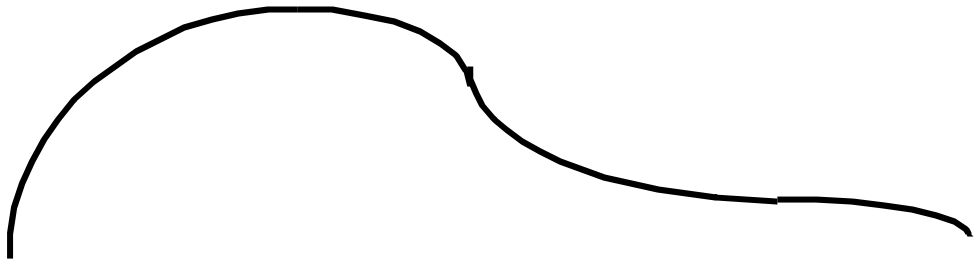


The right part of a unimodal distribution is called the Upper Tail of the distribution while the left part is called the lower tail:

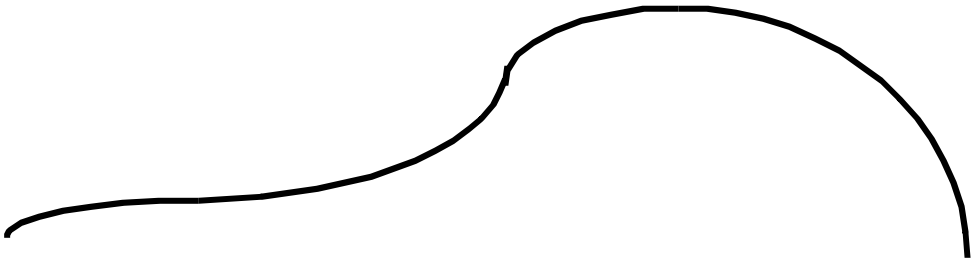


A Unimodal distribution which is not symmetric is called skewed, there are two types of skewness.

Positive Skew: If the upper tail of the distribution stretches out more than the lower tail then the distribution is said to be positively skewed.



Negative Skew: If the Lower tail of the distribution stretches out more than the upper tail then the distribution is said to be negatively skewed.



Section 2.2A Sampling a preview

In Chapter 6, the theory behind Sampling distributions will be covered in detail. But since inferential statistics is based on sampling, and since it will be a while before we reach Chapter 6 perhaps a preview is a good idea. We have seen some of the basic concepts in Chapter 1. Remember the definitions of a **Population** and a **Sample**.

Our aim in Inferential statistics is to make a measurement about some **Characteristic** (property) of the Population but usually the Population is too large for us to perform the required measurement.

So instead we take a **Representative Sample** from the **Population** and measure the value taken by the **Sample Statistic** which corresponds to **Population Characteristic** we are interested in.

So we have two groups of Objects:
The Population and **The Sample**.

And we have two corresponding measures associated with these:
The Population Characteristic and **The Sample Statistic**
An example of this would be

Population = Entire population of Ireland

Sample = A selection of 1000 Irish people chosen at random.

Population Characteristic: We are interested in measuring the Mean (average) Age of the Population of Ireland. We use the Greek letter **m** (pronounced mu) to represent this **Population Mean**.

As already mentioned it would take too much time and effort for us to measure **m.**

Sample Statistic: So Instead we measure the corresponding Mean age for the sample of 1000 people. We use $\bar{x}$ (x bar) to represent this **Sample Mean**.

In this example and <u>in general</u> both **m** and $\bar{x}$ represent the same concept the only difference being that **m** refers to the population and $\bar{x}$ to the sample.

The next two sections will examine which Sample Statistics best measure the Centre of a Population and the Variability of a Population.